[一般論文]

ディジタル・ヒューマニティーズプロジェクト

── 近代公文書自動解読のための基盤的研究 ──

山田　雅之、目加田慶人、長谷川純一、鈴木　哲造
東山　京子、檜山　幸夫、寺沢　憲吾、川嶋　稔夫

1．はじめに

ディジタル・ヒューマニティーズ（Digital Humanities, DH）とは人文科学の調査・研究・教育の方法に情報技術を取り入れ、人文科学の発展を目指す分野である。人類の知的資料のアーカイブ構築やマルチメディアを使った展示方法の研究など具体的な試みは多岐にわたり、文系・理系の枠組みにとらわれない学際的領域に位置づけられる。

我々は日本近代公文書自動解読を目的としたディジタル・ヒューマニティーズプロジェクトを 2015 年度より実施している。2015 年度から 2016 年度までは、研究組織作り、関連研究の調査、研究計画の立案、外部からの研究資金獲得などの活動を実施した。2017 年度からは、近代公文書の手書き文字、翻刻に関するデータの収集とシステム構築に必要となる要素技術の開発を進めている。

人と同じように高度な情報処理を行い、知的に振る舞うシステムを総称して人工知能システムという。このような人工知能システムの機能を実現する方法として、近年、積極的に用いられているものに機械学習がある。

特に、大規模データと深層学習（ディープラーニング）を用いることにより、これまで開発が困難と考えられていた多数の機能が実現されている。本研究プロジェクトでは専門家と同じように近代公文書を解読できる人工知能システムを開発したいと考えている。このシステムの開発には、深層学習のような有望な技術を取り入れていく必要があり、また、近代公文書の特徴を網羅する十分な量のデータや専門家の読み方を取り入れるための新規アイデアが必要である。

以降では、まず、近代公文書を題材とすることの学術的意義を述べ、次に、現在開発を進めようとしているプロトタイプシステムの概要を述べる。また、他の研究プロジェクトで行われた同種の試みをいくつか紹介する。次に、実験を行うために開発した近代公文書データセットについて述べ、その後、そのデータを用いた文字セグメンテーション、文字認識の実験について述べる。

2．研究背景

現在、各行政機関等が保管している戦前期の公文書の多くは、近世古文書の流れを汲む近代古文書のため、古文書解読の知識・経験がないものにとっては、それらを読み解くことは容易ではない。一方、それらは近代の史実が記録された歴史史料であり、広く一般の国民や外国人研究者が利用できるようにすることが望ましい。

手書き文字認識技術については、1968 年に郵便番号自動読み取り機が、1989 年に宛名自動読み取り機がそれぞれ製品化されるなど、初めは対象を制限した形で実用化されてきた。現在では、深層学習を用いた OCR も市販されるようになり、一般の手書き文書に対する認識精度も向上しているはずである。しかしながら、古文書を対象とする場合、後述するように解決すべき課題は多い。それゆえ、複数の研究機関が古文書自動解読を目

図1　台湾総督府文書

指し研究を進めているが、近代公文書をターゲットとしたものは本研究プロジェクトのみである。

この研究プロジェクトを進める上で必要となる近代公文書史料として、台湾総督府文書を活用する。台湾総督府は、日本が台湾を統治していた時代に台湾に設置した行政機関である。また、台湾総督府文書は、明治 8 年（1895 年）から昭和 20 年（1945 年）までのあらゆる種類の公文書（上奏文、法令命令文、内閣文書、各省庁などの関連文書）が原型のままに残された雛形的存在であり、その量は 13,146 簿冊（1 簿冊約 500 ページ×400 字）にのぼる。図 1 にはその内の 7 簿冊を示す。本研究プロジェクトのメンバーは 1982 年から台湾総督府文書研究（現在は台湾総督府文書目録の編纂と同時に進めている台湾総督府文書史料検索データベースの構築）に携わり、その研究で培ってきた古文書解読の知識を本研究プロジェクトに活かすことができる。

3．プロトタイプシステムの概要

古文書は毛筆を使った手書き文書であり、古い字体やくずし字が多用されている。それらを現代使われている字体や楷書（活字）になおし、一般に読める形式することを翻刻という。これは古文書を解読する際には欠かせない作業である。

本研究プロジェクトは自動解読システムを目標とするが、ここで述べるプロトタイプシステムは自動翻刻や翻刻支援を目的としたシステムである。このようなシステムをプロトタイプとする理由は次の二つがある。解読とは意味も含めて読み解くことをいう。コンピュータによる意味理解は、情報学分野の重要テーマの一つであり、また、「意味」をどのように定義するかにより、「意味理解」の定義や意味理解をするための技術的難易度も変わる。本研究プロジェクトにおいては解読をどのように定義するかは、現段階では明確にしていない。また、現段階では意味理解という高次情報処理ではなく、それより低次な文字認識・文書認識のための技術開発をすべきと判断している。このプロトタイプシステムは開発する文字認識・文書認識技術の有効性の検証に利用できる。もう一つの理由は、近代公文書を一般の国民や外国人研究者が利用できるようにする手段として、翻刻結果を提供することが有効であると考えるからである。

3.1．システムの機能

大量の文書を翻刻する場合、人件費や作業効率を考慮した翻刻体制をとる。図 2 は翻刻作業に携わる人員を知識レベルに応じて、初級者、中級者、上級者に分けて分業する体制を示す。人件費の安い複数の初級者が一次的な翻刻を行い、中級者、上級者は翻刻結果を点検し、翻刻誤りを修正する二次的翻刻を行う。システムはこのような翻刻作業を支援するため、次の

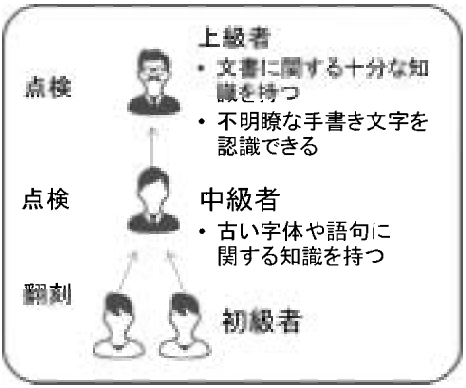

図 2　翻刻作業体制の例

三つの機能を提供するものとする。

⑴　翻刻支援機能

初級者が文書（ページ）内の一部の範囲を指定し、システムがその範囲内に含まれる手書き文字の翻刻候補を提示する機能

⑵　点検支援機能

タイプミスや翻刻誤りを検出する機能

⑶　自動翻刻機能

初級者や中級者の代理となり自動翻刻する機能

3.2. システムで用いる手書き文字・文書認識手法と事前に必要なデータ

3.2.1. 手書き文字認識

文字認識とは平面上に分布するドットパターンあるいは濃淡パターンから、そのパターンが表す文字を推定することである。手書き文字は筆記者の違いによりパターンが変化するため、活字に比べ認識は難しい。

本研究プロジェクトでは、前述のシステムを開発するために、お互いを補い合う二つの文字認識手法を用いることを検討している。一つは機械学

習に基づく文字認識、もう一つは文字スポッティングである（後者はより一般的に、ワードスポッティングと呼ぶことが多い)。

文字認識は手書き文字（パターン）がどの字種（カテゴリ）に属すかを推定する識別問題として扱うことができる。この問題をコンピュータで解くための代表的な手法に機械学習がある。カテゴリ別に分類されたパターンの集合を事前に用意し、これを利用して、同じカテゴリのパターンに共通する特徴や異なるカテゴリのパターンを差別化する特徴を見つけ出し、それらの特徴に基づく識別ルールを求める。このとき、ルールを徐々に一般化していき、汎用性の高いものにすることを情報学の分野では学習と呼ぶ。

文字スポッティングも、手書き文字認識のためによく用いられる手法の一つである[1]。これは認識したい手書き文字とデータベースに格納されている手書きの文字を、文字の形に基づいて照合する手法である。

両手法を比較すると、機械学習に基づく文字認識は、筆記者の多様性に対する頑健性が高く、また、文字認識の速度も速い。その反面、機械学習に基づく文字認識は、認識しようとする手書き文字の範囲の切出し（文字セグメンテーション）が正確に行われないと認識精度が低くなるが、文字スポッティングは文字セグメンテーションの正確さに依存しない。

3.2.2. 手書き文書認識

一つ一つの文字を個別に認識することを文字認識と呼ぶのに対し、パラグラフやページの単位でその中の全ての文字を一括して認識することを、文書認識と呼ぶ。コンピュータによる文書認識の典型的手順は次のようなものである。まず、文書から行を検出し、次に、各行を文字単位に分割する。これにより文字セグメンテーションが完了する。この後、個々の文字の認識を行うが、認識結果として出力する字種を一つに絞るのではなく、ある程度信頼性のある複数の候補を出力する。どの候補が適切であるかは、

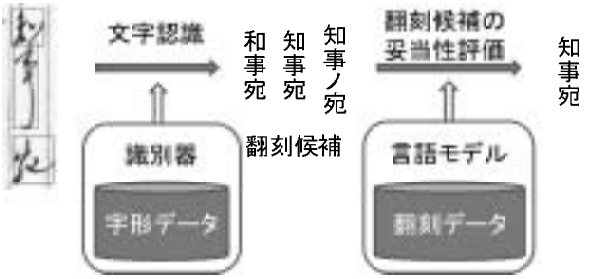

図 3　文書認識処理の例

前後の文字の認識結果と言語レベルの知識を用いて判定する。ここで、言語レベルの知識とは、例えば、単語辞書や、膨大な量の文書データを解析して得られる統計的情報（言語モデルと呼ぶ）のことである。なお、1989年に実用化された宛名自動読み取り技術では、地名や人名の辞書を用いている。図 3 は一連の処理過程を示す。

3.2.3. 必要となるデータ

字種ごとに分類された手書き文字のデータを字形データと呼ぶ。文字認識を行うためには、十分な量の字形データが必要である。本研究プロジェクトでは、台湾総督府文書をページごとにスキャンして得られた画像を利用して字形データを作成する。

また、文書認識を行うためには、近代公文書に関する言語レベルの知識が必要である。この知識を機械的に構築するためには、統計的分析が可能な形式である文書データが十分な量必要である。本研究プロジェクトでは、このような文書データを、台湾総督府文書を翻刻して作成する。ここでは、台湾総督府文書を翻刻したものを翻刻データと呼ぶ。また、台湾総督府文書目録データベース[2]や以降で述べる「古語的難解語句・慣用表現例リスト」を利用して、近代公文書によく表れる単語や慣用表現に関する辞書も作成する。

4．関連研究

2000 年代以降、手書き文字認識の精度は急速に向上している。この理由として、畳込みニューラルネットワーク[3]という、パターンの変化に頑健な認識技術が用いられるようになったこと、深層学習フレームワークや手書き文字認識・文書認識の実験に必要なデータセットが公開され、この分野に参入しやすくなったことなどがあげられる。中国科学院自動化研究所（CASIA）が 2010 年に公開した中国語データベースには、筆記者 1,020 人、7,185 字種、3,721,874 文字の字形データと筆記者 1,019 人、5,091 ページ、52,230 行、2,703 字種、1,349,414 文字の文書データが含まれており（図 4 参照)、近年、これを用いた研究成果が多数報告されている。株式会社富士通研究所のプレスリリース[4]では、文字列（行単位）認識において 96.3%の精度を達成したと報告されている。

歴史的文書の翻刻にコンピュータを活用する試みも国内外で行われている。ここでは、日本の古文書を対象とした 3 種類のデータセットと、それらデータセットを用いた翻刻支援システムや文字認識実験を紹介する。

古文書翻刻支援システム開発プロジェクト（HCR プロジェクト)[5][6]では、2001 年時点で、「伏見屋善兵衛文書」(文化から慶応年間にいたる各種の証文類約 1300）から約 243,000 文字、「宗門改帳」から 16 字種(主に漢数字)、7,866 文字、東京堂出版「毛筆版くずし字解読辞典」から 25,202 文字の手書き文字を収集しデータベース化している。ニューラルネットワークを用いた 16 字種の文字認識実験では 97.05%の精度を達成している[7]。

奈良文化財研究所・東京大学史料編纂所の木簡・くずし字解読システムには、古代から近世の様々な古文書から切り出した約 30,000 の字形画像が登録されている。このシステムに 1 文字分の画像を入力すると、類似す

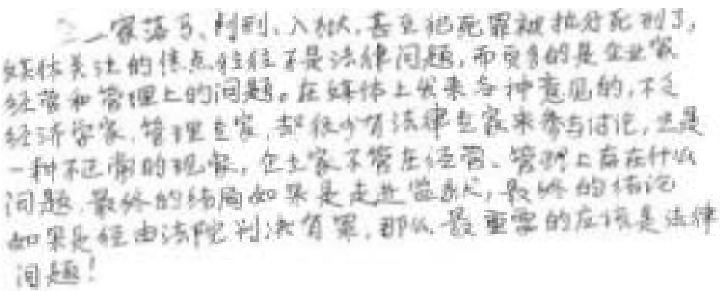

図 4　CASIA 中国語データベースに含まれる手書き文書の例

る字形画像とその文字コードを検索できる[8]。

国立情報学研究所と国文学研究資料館は、主に料理レシピを扱った江戸時代の古典籍をもとにして日本古典籍字形データセットを作成し、2017 年 12 月現在、約 40 万文字のデータを公開している。山本らはこのデータを用いて、ワードスポッティングによる文字認識実験を実施し、80%以上の精度が得られたと報告している[9]。早坂らはこのデータセットを用いた機械学習が、他の文書の文字認識に有効かどうかを検証する実験を行っている。機械学習には畳込みニューラルネットワークを使った深層学習の手法を用いている。「平家物語」、「源氏物語」に現れる平仮名を対象とした場合、それぞれ 91%、95%の認識精度であったと報告している[10]。

ここで紹介した古文書のデータセットはその文書の書かれた年代や、その文書が扱っている内容が異なる。そのため、データセットに含まれる手書き文字の特徴や文体の特徴なども異なる。文字認識・文書認識の技術を開発する際には、実験対象となるデータセットについて、事前にそれらの特徴を理解しておく必要がある。

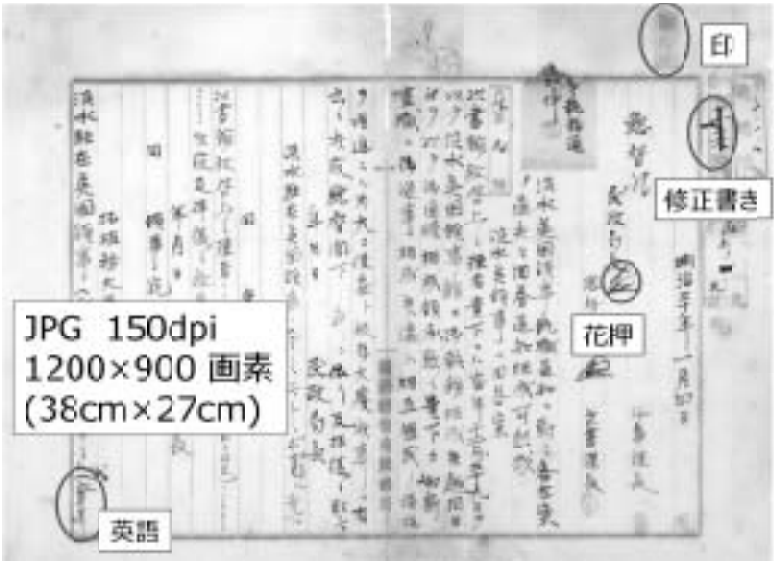

図 5　台湾総督府文書の 1 枚の画像

5．台湾総督府文書における手書き文字の特徴

本研究プロジェクトでは、台湾総督府文書を利用して、文字認識・文書認識の実験に用いるデータセットを作成する。このため、台湾総督府文書がどのように書かれているのかを事前に把握しておく必要がある。

図 5 は台湾総督府文書の 1 枚（1 ページ）をスキャナーで読み込み得られた画像である。これを原画像と呼び、字形データを作成する際には、個々の手書き文字は原画像から採取する。また、原画像は文書認識実験の対象ともなる。図 5 の原画像を見ると、毛筆で書かれた漢字や片仮名の手書き文字のみでなく、活字、英字、記号、罫線、印、修正書き、花押、汚れ、破れなどがあることも確認できる。

図 6 は手書きスタイルの多様性を示す。筆記者や年代などの違いから、

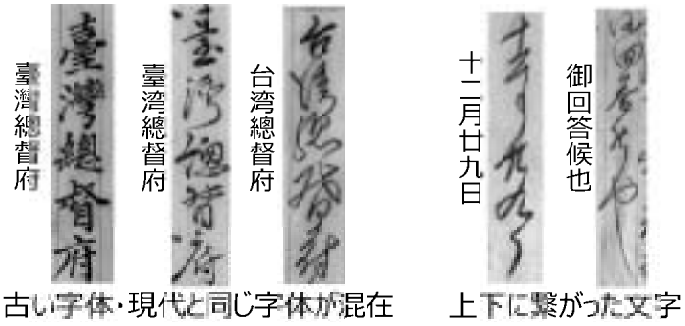

図 6　多様な手書きスタイル

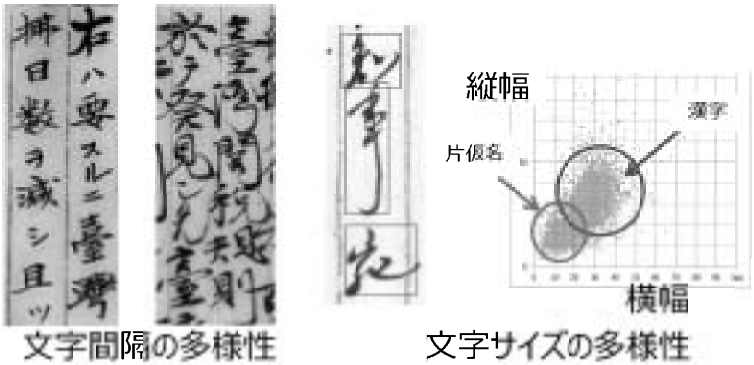

図 7　多様な文字間隔・文字サイズ

「臺」/「台」のような古い字体/現代の字体が使われている。また、上下の文字がつながっているなど、縦書き特有の特徴も確認できる。

図 7 は文字間隔や文字サイズの多様性を示す。文字と文字の間にスペースを入れたものや、逆に、詰めて書いたものも確認できる。また、縦長、横長の文字があり、文字のサイズも多様である。図 7 右のグラフは文字サイズの分布を示す。片仮名は漢字に比べ小さく書かれる傾向がある。

このように、人にとって読みやすいものから、読みにくいものまで、様々な難易度の手書き文字・文書があるが、この難易度はコンピュータによる

文字認識・文書認識の精度にも影響する。そして、システムの精度を高めるためには難易度の高い手書き文字・文書を正しく認識する必要がある。

6．近代公文書データセットの開発

我々は技術開発や各種実験を円滑に行うため、データセットの開発に取り組んでいる[11]。このデータセットは、台湾総督府文書の原画像、翻刻データ、字形データ、古語的難解語句・慣用表現例リストから構成される。

6.1．原画像

原画像は国士館臺灣文献館から提供されている。原画像の画像フォーマットはJPGであり、ファイル名は簿冊番号、ページ番号が分かる形式となっている。例えば「000001310030014.jpg」は、簿冊番号131、ページ番号14であることを表す。現在、131番、148番、213番、2280番、5935番, 10109番の簿冊に含まれる計2,466ページ分の原画像をデータセットに収めている。

6.2．翻刻データ

翻刻は大学院生、学部生の協力を得ながら、先に示した翻刻体制によって進められている。2017年12月時点で、年代の異なる上記6簿冊の中からサンプリングした853ページ分の翻刻データを作成できている。但し、手書き文字を翻刻し、活字や印は翻刻を省略している場合もある。図8に翻刻データの例を示す。行、削除部分、挿入部分に関して原画像と翻刻データが対応付けできる形式となっている。また、各行が前の行に続く行であるか否かが分かるように、続く行の先頭には「＋」記号を付けている。

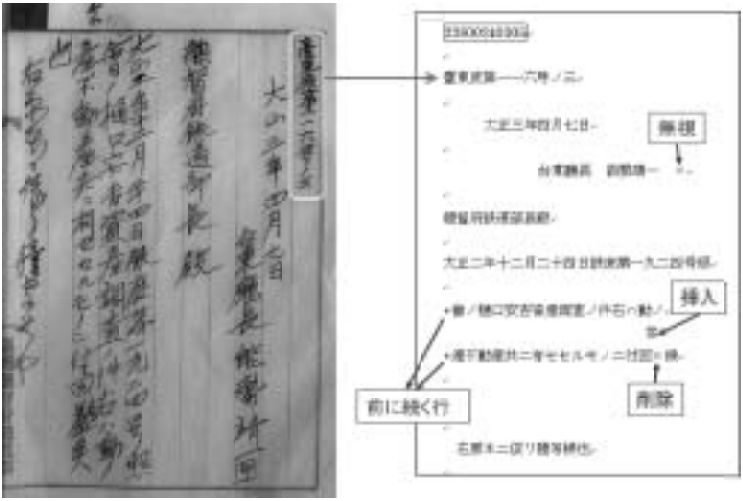

図 8 翻刻データの例

6.3. 字形データ

字形データは個々の手書き文字の原画像における位置 (x, y)、サイズ (w, h) およびその手書き文字が表す字種を記録したものであり、次のような csv 形式としている。

原画像ファイル名, x, y, w, h, 字種

図 9 は字形データを作成する流れを示す。字形データを作成するには、個々の文字の原画像中における位置、サイズを求める必要がある。これらを得るために画像処理よる文字セグメンテーションを行い、その結果が不適切であれば手作業で修正している。その後、翻刻データの文字と、画像中の文字セグメントとを対応付けてデータを作成している。2017 年 12 月現在、約 176,000 文字、2600 字種の字形データが得られている。文字サイズは平均 21×23 画素であった。176,000 文字の内、約 23%は片仮名であり、最も文字数の多い片仮名は「ニ」(4,648 文字) である。台湾総督府文書は助詞や送り仮名として片仮名が主に使われているため、ページ単

図 9　字形データ作成の流れ

図 10　字形データに含まれる「ﾄﾓ」「灣」の手書き文字の例

表 1 古語的難解語句・慣用表現例リスト(一部の例)

況（いわんや)、抑（そもそも)、忝（かたじけなし)、偖（さて)、抔（など)、并ニ(ならびに)、迚モ（とても)、加之（しかのみならず)、都而（すべて)、決而（けっして)、乍去（さりながら)、乍恐（おそれながら)、有之（これあり)、依之（これにより)、斯様（かよう)、向後（きょうこう)、穴賢（あなかしこ)、幾許（いくばく)、今以（いまもって）事由（ことのよし)、事之外（ことのほか)、事候間（ことに・そうろう・あいだ)、悪敷者（あしきもの)、無覚束（おぼつかなし)、一件（いっけん)、如仰（おおせの・ごとし)、如件（くだんのごとし)、仍如件（よって・くだんのごとし)、如此(かくの・ごとし／ごとく)、如故（もとの・ごとし)、如何（いかん／いかに)、如何様(いかよう)、如何共（いかん・とも)、如来（にょらい)、如意（にょい)、如是（かくの・ごとく）

位、文章単位でデータを収集すると、字種の多い漢字と比べ、個々の片仮名の頻度が大きくなる。なお、最も文字数の多い漢字は「十」(2,924 文字) であった。図 10 には収集した手書き文字の例として、合字「ﾄﾓ」(20 文字) と「灣」(435 文字) の一部を示す。

6.4. 古語的難解語句・慣用表現例リスト

これは近代公文書に用いられる古語的な語句・慣用表現をリストしたものである。これは、初級者の翻刻を助けるものであるが、システムによる文書認識の精度を高めるためにも利用できる。現在、計 342 の語句・慣用表現を収めている。表 1 にその一部を示す。

7. 実験

ここでは、文字セグメンテーションと文字認識の実験について述べる。両者ともに我々が開発したデータセットを実験試料としたものである。

7.1. 文字セグメンテーションの実験

文字セグメンテーションとは、文書中の文字列を個々の文字に切り分け

ることである。手書き文書の場合、個々の文字のサイズや文字間隔は一定ではないため、手書き文字認識と同様、文字セグメンテーションも容易ではない。本実験では、原画像を入力とし、文字セグメンテーションの結果を出力するまでの処理を、背景除去、行検出、セグメント統合に分けて順次行う。

7.1.1. 背景除去処理

これは、文字のみを残し、それ以外を除去した画像を生成する処理である。画像は格子状に並ぶ画素で構成されており、この処理では、各画素が文字の一部であるかどうかを判定する。その判定には画素の色情報を用いるが、判定しようとする一つの画素（注目画素と呼ぶ）の色情報のみでは、文字の一部であるかどうかを正しく判定することは難しい。なぜなら、文字を構成する画素の色は、人には黒やグレーに見えても、RGB の数値上ではそうではない。また、罫線や何も書かれていない紙面の部分の画素が、文字の画素と近い色である場合もある。そこで本実験では、注目画素のみでなく、その周辺の画素の色情報も用いて、注目画素が文字の一部であるかどうかを判定する。また、判定ルールの獲得には、全層畳込みネットワークモデル[12]を使った深層学習を用いる。判定ルールの学習のためには学習用のサンプル画像を用意する必要がある。このために、文字・罫線の明るさ・色が互いに異なる原画像を選び、その画像の一部を 256×256 の大きさに切り出した計 526 枚の画像と、それらに対して正しく背景除去を行って得られる画像を作成し、学習をさせた。

図 11b の画像が背景除去処理の結果である。図 11b では一つの原画像に対する処理結果の一部分を示しているが、画像全体に対する精度は適合率 80%、再現率 89%であった。ここで適合率は、文字の画素と判定したもののうち、真に文字の画素であるものの割合であり、再現率は、真に文字の画素であるもののうち、文字の画素と正しく判定されたものの割合で

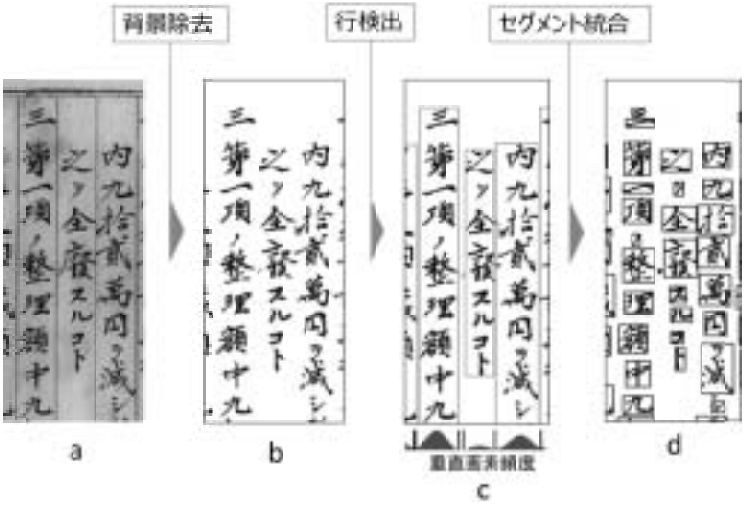

図 11　原画像を入力とし、文字セグメンテーション結果を出力するまでの処理手順

ある。複数の原画像に対する背景除去処理の結果を見ると、文字の形状を維持しつつ、文字以外が除去されていることを確認できた。

7.1.2. 行検出処理

行検出は、まず、垂直方向に文字の画素を数え、次に、その総数が小さい値となる水平位置を求め、その位置を隣接する二つの行の区切りとする。図 11c の画像は、行を検出した結果であり、その下には垂直方向の文字画素の頻度を示す。この方法は行が垂直であることを前提としており、そうでない画像に対しては、予め回転補正を行う。

7.1.3. セグメント統合処理

この処理は、まず、文字の画素が連結している成分（セグメント）を見つける。図 12 左側にその様子を示す。図 12 左上の片仮名「ル」は二つのセグメントで構成されている。次に、同じ行内にあり横方向に隣接するセ

グメントを統合することを繰り返す。図 12 右側と図 11d はセグメント統合処理の結果である。この方法は、文字の画素が連結しているセグメントを最小単位としている。そのため、図 11d の一行目の「ヲ」と「減」は一つのセグメントになっている。また、縦方向に隣接するセグメントの統合は行わないため、図 11d の 3 行目の「三」が三つのセグメントに分かれたままである。図 11 の原画像全体に対する文字セグメンテーション精度は 85%であったが、この方法では、精度が手書きスタイルに影響されやすい。安定した精度を得るためには、より頑健な文字セグメンテーション手法が必要である。

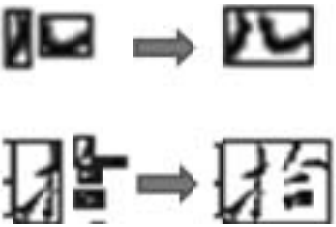

図 12　セグメントの統合の例

7.2. 文字認識の実験

この実験では、字形データから得られる手書き文字の画像と、畳込みニューラルネットワーク（Convolutional Neural Network, CNN）を用いて、台湾総督府文書における手書き文字の認識を行い、その精度を計測する。CNN の機能や学習の原理は次のようなものである。CNN へは一定の大きさの手書き文字画像を入力する。CNN の出力は入力された画像がどの字種に該当するかを表す確率であり、これを確信度とも呼ぶ。例えば字種が 3 種類 A、B、C のみの場合、出力されるのは（0.7, 0.1, 0.2）の様な数値列である。これは A、B、C である確信度がそれぞれ 0.7, 0.1, 0.2 であることを表す。この例の場合、最も確信度の大きい A を認識結果の第 1 候補とする。CNN の内部では、入力画像のデータに基づいて確信度を計算する。学習のフェーズでは、字種 A の手書き文字のサンプルを十分用意し、それらが入力されたときは確信度の列（1.0, 0.0, 0.0）が出力されるように、内部計算で用いるパラメータを調節する。同様に、字種 B、C

についても、それぞれの手書き文字のサンプルを十分用意し、それらが入力されたとき（0.0, 1.0, 0.0）、（0.0, 0.0, 1.0）が出力されるように内部計算のパラメータを調節する。このときパラメータは少しずつ更新し、これらの手順を繰り返しながら、全てのサンプルに対して正しい確信度が出力されるようにする。テストフェーズでは、学習を済ませた CNN の識別能力を評価するため、学習のときとは異なるサンプルを入力し、その認識結果をみていく。

7.2.1. 実験方法

CNN へ入力する手書き文字画像のサイズは 48×48 画素とする。台湾総督府文書の原画像から得られる手書き文字のサイズは一定ではなく、平均は 21×23 画素である。そのため、手書き文字画像を 48×48 画素にリサイズし、入力することになる。

学習フェーズとテストフェーズで用いる字形データの総数を K とし、この実験では、K＝64,655、K＝15,365 の二とおりの場合について、それぞれ精度を計測した。K 個の字形データは次のようにして、学習フェーズ用とテストフェーズ用に分ける。まず、K 個のデータを均等に 5 グループに分け、そのうち 3 グループを学習フェーズ用に、1 グループをテストフェーズ用にする。学習フェーズとテストフェーズに割り振るグループを組み替えながら、計 5 回の学習・テストを行い、各テストフェーズにおける認識率を計測し、5 回分の平均値を全体の精度とする。

7.2.2. 結果

図 13 には、精度をグラフ化したものを示す。縦軸は第 N 候補までに正しい字種が含まれる割合、横軸は N である。データ数 K＝64,655 の場合、第 1 候補が正しい字種である割合は 75%であり、第 5 候補までに正しい字種が含まれる割合は 88%であった。また、K＝15,365 の場合に比べ、

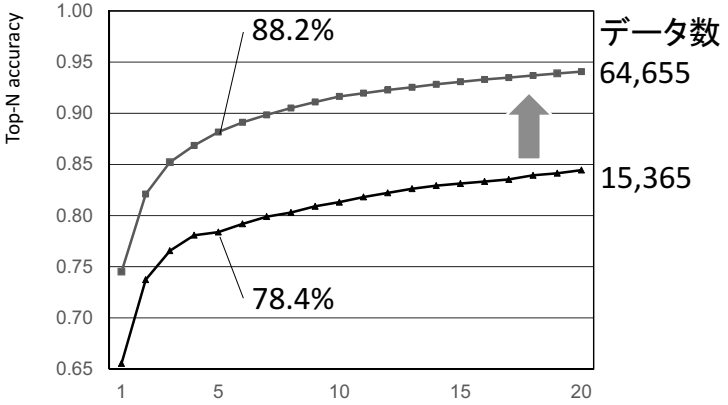

図 13　縦軸は第 N 候補までに正しい字種が含まれる割合、横軸は N

精度が 10%程度上がっている。これらの結果から学習のデータ数を増やせば、さらに精度が向上することが期待でき、有意な翻刻支援[13]が行える可能性がある。

8．考察

台湾総督府文書は近代公文書特有の文体であり、また、その手書き文字は認識が容易なものから、難しいものまであり、多様性の幅が広い。これらは、他の研究機関が対象としている古文書と異なる特色であり、このような文書を対象とすることが本研究プロジェクトの一つの大きな特色である。また、開発しているデータセットは日本語文字認識の研究分野に寄与するものとなる。したがって、今後もデータセットの開発を進め、量を増やしていく予定である。しかしながら、データセットの開発は人力によるところが大きく、労力と時間を要する。また、現在の作成方法では、ページ単位の翻刻データをまず作り、その後、それに対応する字形データを作っている。この手順では、字種ごとの字形データ数に偏りができる。このた

め、字形データ数の少ない字種については、別途データを収集するなどして偏りを減らしていく必要がある。

文字認識の実験については、利用する字形データ数を増やすことにより認識精度が向上したことを述べた。字形データ数を増やす方法として、台湾総督文書から採取する方法の他に、採取済みの手書き文字を疑似的に変形し、それを新たな字形とする方法もある。また、他のデータセットを併用する方法もある。古典籍字形データセットに含まれる手書き文字の特徴は本データセットのものと類似するところがあり、その一方で、字種ごとの字形データ数の分布が本データセットと異なる。したがって、古典籍字形データセットの併用も有効であると思われる。

文書認識においては、文書構造や意味に関して整合性ある認識結果を出力する必要がある。その一つの手順として、本稿では、文字セグメンテーションや文字認識など低次の処理を行い、その後、整合性をチェックする方法を述べた。しかしながら、これは古文書専門家の読み方とは大きく異なる。古文書専門家は、文書構造や意味を把握しながら読み進める大域的な読みと、個々の手書き文字を読み取っていく局所的な読みをシームレスに繰り返しながら、文書認識を行っていると考えられるからである。今後は、この差異を小さくする方法についても検討していきたい。

9. おわりに

本稿では、ディジタル・ヒューマニティーズプロジェクトの特色、プロトタイプシステムの概要、近代公文書データセット、文字セグメンテーションの実験、文字認識の実験について述べた。実験については、主に、中京大学のプロジェクトメンバーが行ったものを取り上げたため、公立はこだて未来大学に所属するメンバーが行っている実験については、別途、報告したい。なお、中京大学と公立はこだて未来大学のメンバーが会した研究

会を既に 4 回開催しており、今後も互いの進捗状況を報告し、議論しながらプロジェクトを推進していく。

謝辞

本研究は第 45 回三菱財団人文科学助成、JSPS 科研費 JP17K03049 の助成を受けた。

参考文献

[1] 固有空間法と DTW による古文書ワードスポッティング、寺沢憲吾、長崎健、川嶋稔夫、電子情報通信学会論文誌. D、情報・システム J89-D (8)、1829-1839、2006.

[2] 台湾総督府文書目録データベース、中京大学社会科学研究所台湾史研究センター、2010.

[3] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. of the IEEE, 2278-2324, 1998.

[4] 株式会社富士通研究所 PRESS RELEASE (2016 年 11 月 8 日付)、人工知能モデルを活用した高精度の手書き文字列認識技術を開発、2016.

[5] 山田奨治、笠谷和比古、加藤寧、並木美太郎、川口洋、原正一郎、石谷康人、小島正美、梅田三千雄、山本和彦、柴山守、古文書翻刻支援システム開発 (HCR) プロジェクト報告 (2)、情報処理学会研究報告、人文科学とコンピュータ (CH) 2001 (51 (2001-CH-050))、9-16、2001.

[6] 山田奨治、柴山守、科学研究費補助金研究成果報告書、古文書翻刻支援システムの研究 (2)、2001.

[7] 柴山守、科学研究費補助金研究成果報告書、古文書文字認識システムの高精度化に関す研究、2005.

[8] 耒代誠仁、井上幸、高田祐一、方国花、馬場基、渡辺晃宏、井上聡木、簡およびくずし字のデジタルアーカイブを文字画像で検索するサービスの実装、じんもんこん 2016 論文集、19-24、2016.

[9] 山本純子、大澤留次郎、古典籍翻刻の省力化：くずし字を含む新方式 OCR 技術の開発、情報管理、58、11、819-827、2016.

[10] 早坂太一、大野亙、加藤弓枝、山本和明、深層学習による変体仮名翻刻アプリケーション開発の試み、第 31 回人工知能学会全国大会、3Q1-2in1、2017.

[11] 釜谷勇輝、山田雅之、目加田慶人、長谷川純一、檜山幸夫、東山京子、中貴俊、宮崎慎也、寺沢憲吾、川嶋稔夫、近代公文書自動解読のための手書き字形データセット構築、平成 29 年度電気・電子・情報関係学会東海支部連合大会、B5-9、2017.

[12] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, In CVPR, 2015.

[13] 楊宗哲、道満恵介、山田雅之、目加田慶人、畳込みニューラルネットワークを用いた日本近代公文書文字認識、平成 29 年度電気・電子・情報関係学会東海支部連合大会、F2-1、2017.